JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 55, 239–252 (1976)

Dynamic Programming of Expectation and Variance*

GISBERT QUELLE

Institut für Mathematische Stochastik der Universität Hamburg, Bundesrepublik Deutschland D2000 Hamburg 13, Rothenbaumchaussee 45

Submitted by M. Aoki

1. INTRODUCTION

The decision model DM of Hinderer [1, 2] is extended to treat the minimization of variance in the set of those plans, which guarantee the maximal expected profit.

The well-known semi-Markovian decision model with discount (cf. Jewell [6], Howard [3], Osaki and Mine [8]) is formulated as such a decision model SMDM, and thus it is to be seen that there exists a deterministic stationary Markovian plan, depending only on states, which is expectation-optimal in the set of all plans, generally depending on all past states, actions and transition times between the states. This plan can be determined by a linear program or by value and plan iteration.

Then we look for the minimization of variance in the set of expectation-optimal plans. This problem, again, is formulated as a semi-Markovian decision model SMDMV, and thus results analogous to those mentioned above are obtained. Meanwhile similar questions were treated by Mandl [7] and Jaquette [4, 5], but with other tools and in a different context. [7] deals with absorbing and recurrent Markovian models, whereas the models and policies in [4, 5] are Markovian under discounting. Both papers use exponential transforms.

1. THE GENERAL DECISION MODEL DM

In this part we give results of Hinderer [1, 2] with a simplified finite action space, and in slight extension we add Lemma 1.4 and Theorem 1.12.

DEFINITION 1.1. We call a tupel $((S, \mathfrak{S}), (A, \mathfrak{A}), (D_n), (q_n), (r_n))$ a *decision model* DM, if the following properties are given:

(a) $(S, \mathfrak{S})$ is the standard Borel *state space* with it's σ–algebra.

* This paper is an extract of the author's doctoral dissertation [9].

Copyright © 1976 by Academic Press, Inc.
All rights of reproduction in any form reserved.

(b) A is the finite *action space*, $\mathfrak{A} := \mathfrak{P}(A)$ with $\mathfrak{P}(A)$ as the set of all subsets of A (by ":=" we mean the defining equality, and $\varnothing$ will denote the empty set).

(c) For all $n \in \mathbb{N}$ (the set of natural numbers) we introduce $\overline{H}_n := S \times A \times \cdots \times S \times A \times S$ ($2n - 1$ factors) as the set of *histories* $h_n := (s_1, a_1, \ldots, s_{n-1}, a_{n-1}, s_n)$, $s_k \in S$, $a_k \in A$, $k = 1, 2, \ldots$, which can possibly have occured up to time n. The σ-algebra of $\overline{H}_n$ is $\overline{\mathfrak{H}}_n := \mathfrak{S} \otimes \mathfrak{P}(A) \otimes \mathfrak{S} \otimes \cdots \otimes \mathfrak{S} \otimes \mathfrak{P}(A) \otimes \mathfrak{S}$.

(d) The *range* $D := (D_n)$ is a set of mappings $D_n : H_n \to \mathfrak{P}(A) \setminus \varnothing$, defined on some $H_n \subset \overline{H}_n$, $n \in \mathbb{N}$. D_n and H_n are related by $H_1 = S$ and $H_{n+1} := \{(h_n, a, s) \in \overline{H}_{n+1} : h_n \in H_n, \ a \in D_n(h_n), \ s \in S\}$, $n > 1$. For a present $h_n \in H_n$, $D_n(h_n)$ is the set of *admissible actions*, and H_n may be interpreted as the permitted part of the history.

(e) q_0 is a probability on $(S, \mathfrak{S})$, the *initial distribution*. For $n \in \mathbb{N}$, q_n is a transition probability from $(\overline{H}_n \times A, \overline{\mathfrak{H}}_n \otimes \mathfrak{P}(A))$ to $(S, \mathfrak{S})$, the *transition law* between time n and $n + 1$.

(f) $(r_n, n \in \mathbb{N})$ is a sequence of extended real-valued $\overline{\mathfrak{H}}_n \otimes \mathfrak{P}(A) \otimes \mathfrak{S}$-measurable functions. $r_n(h_n, a, s)$ evaluates the *reward* resulting from a transition h_n to s under action a.

DEFINITION 1.2. A *plan* π is a sequence (π_n) of transition probabilities from $(\overline{H}_n, \overline{\mathfrak{H}}_n)$ to $(A, \mathfrak{P}(A))$. The plan π is said to be *deterministic, if* the π_n are concentrated on points $f_n(h_n) \in A$. This π will be denoted by $f = (f_n)$.

According to a theorem of C. Ionescu-Tulcea, to a given plan π there exists a unique probability measure Q_π on the space $(\overline{H}, \overline{\mathfrak{H}}) := (S \times A \times S \times \cdots, \mathfrak{S} \otimes \mathfrak{P}(A) \otimes \mathfrak{S} \otimes \cdots)$. The structure of Q_π-integrals can be seen in [1]. For convenience, we denote $Q_\pi(B \times A \times S \times \cdots)$ by $P_{n\pi}(B)$, $B \in \overline{\mathfrak{H}}_n$, and using a simplified symbolism for integration, we can write $Q_\pi = q_0\pi_1 q_1 \pi_2 q_2 \cdots$, and $P_{n\pi} = q_0\pi_1 q_1 \pi_2 \cdots \pi_{n-1} q_{n-1}$. In the same way, we symbolize the transition probability $(\pi_n q_n \pi_{n+1} q_{n+1} \cdots)(h_n ; \cdot)$ by $Q_{n\pi}(h_n)$.

DEFINITION 1.3. A plan π is called *almost sure admissible* under range D(*a.s. admissible*), if $\pi_n(h_n ; D_n(h_n)) = 1$ for $P_{n\pi}$ — a.a. $h_n \in H_n$, $n \in \mathbb{N}$. Δ^D is the set of a.s. admissible plans. A plan π is called *sure admissible* under range D (*admissible*), if $\pi_n(h_n ; D_n(h_n)) = 1$ for all $h_n \in H_n$, $n \in \mathbb{N}$. Δ is the set of admissible plans.

Obviously we have $\Delta \subset \Delta^D$. It is easy to see from the theorem of Ionescu-Tulcea, that $P_{n\pi}(H_n) = 1$ for all $n \in \mathbb{N}$, $\pi \in \Delta^D$.

LEMMA 1.4. *For $\pi \in \Delta^D$ there is a $\sigma \in \Delta$ with $Q_\pi = Q_\sigma$.*

Proof. For $\pi \in \Delta^D$, $n \in \mathbb{N}$, let $H_n{}^\pi \in \mathfrak{H}_n$ be the set with $P_{n\pi}(H_n{}^\pi) = 1$, that fulfills $\pi_n(h_n\,;\, D_n(h_n)) = 1$ for $h_n \in H_n{}^\pi$. To $n \in \mathbb{N}$, $h_n \in H_n$, choose $a_n(h_n) \in D_n(h_n)$. With δ_a denoting the point-mass in a, we define the plan σ by

$$\sigma_n(h_n;\,\cdot) := \begin{cases} \delta_{a_n(h_n)}(\cdot) & \text{for } h_n \in H_n \backslash H_n{}^\pi, \\ \pi_n(h_n;\,\cdot) & \text{elsewhere.} \end{cases}$$

Using the theorem of Ionescu-Tulcea we confirm the assertion by induction.

Let χ_n be the projection from $\bar{H}$ to $\bar{H}_n$, the history at time n. If they exist, let be: $R_n(h) := \sum_{k=n}^{\infty} r_k(h_k, a_k, s_{k+1})$, $R(h) := R_1(h)$, $V_\pi := E_\pi R := \int R\, dQ_\pi$, $V_{n\pi}(h_n) := E_\pi[R_n \mid \chi_n = h_n] := \int R_n\, dQ_{n\pi}(h_n)$, $n \in \mathbb{N}$, $\pi \in \Delta^D$.

DEFINITION 1.5. If $V_{\pi^*} := \sup_{\pi \in \Delta^D} V_\pi$ exists, we call the plan $\pi^* \in \Delta^D$ *expectation-optimal* under range D (*optimal*).

COROLLARY 1.6. *There exists an optimal plan in Δ^D if and only if there exists one in Δ, in which case for $\pi \in \Delta^D$ there is some $\sigma \in \Delta$ with $E_\pi R = E_\sigma R$.*

Now we are justified in restricting our attention to Δ. Hinderer [2, Lemma 2.1] permits the definition of $W_{n+} := \sup_{\pi \in \Delta} E_\pi[R_{n+} \mid \chi_n = \cdot]$ with $R_{n+} := \sum_{k=n}^{\infty} r_k{}^+$, $r_n{}^+ := \max\{r_n, 0\}$, and $T_n v := \sup_{a \in D_n(\cdot)} \int q_n(\cdot, a;\, ds)\, v(\cdot, a, s)$, $T_{nk}\, v := T_n T_{n+1} \cdots T_{n+k-1} v$, n, $k \in \mathbb{N}$, for any integrabel function v. We introduce the important.

ASSUMPTION 1.7. GV: $\int W_{1+}\, dq_0 < \infty$ and $W_{n+} < \infty$, $n \in \mathbb{N}$.
C$^+$: $\lim_k T_{nk} W_{n+k,+} = 0$, $n \in \mathbb{N}$.

(GV $\equiv$ general assumption on values, $C^+ \equiv$ convergence of positive rests). Henceforth GV will be presumed as valid, even if not mentioned explicitly.

Reference [2, Lemma 3.1] affirms the existence of $V_{n\pi}(h_n)$, and we introduce $V_n := \sup_{\pi \in \Delta} V_{n\pi}$, $V := \sup_{\pi \in \Delta} V_\pi$, $n \in \mathbb{N}$. [2, Satz 3.2], [2, Satz 3.3] and Rieder [10, Satz 4.1] prove the following results:

LEMMA 1.8.

(a) *V_n is universally measurable, and also universally measurable in all coordinates and couples of coordinates.*

(b) *(V_n) satisfies the optimality equation OE*, i.e.,

$$V_n(\cdot) = \sup_{a \in D_n(\cdot)} \int q_n(\cdot, a;\, ds)\, [r_n(\cdot, a, s) + V_{n+1}(\cdot, a, s)].$$

LEMMA 1.9 [1, Theorem 15.2; 2, Satz 4.1]. *Let C^+ be given. Then for $\pi \in \Delta$ there is a deterministic $f \in \Delta$ with $V_{nf} \geqslant V_{n\pi}$, $n \in \mathbb{N}$.*

LEMMA 1.10 [1, Theorem 17.1]. *If* $-\infty < V < \infty$, *the following three statements about* $\pi \in \Delta$ *are equivalent*:

(a) π *is optimal*

(b) $V_{1\pi} = V_1 q_0$ — a.s.

(c) $V_{n\pi} = V_n P_{n\pi}$ — a.s., $n \in \mathbb{N}$.

DEFINITION 1.11. For $n \in \mathbb{N}$, $h_n \in H_n$ and $a \in D_n(h_n)$ define $(L_n v)(h_n, a) := \int q_n(h_n, a; ds)\ [r_n(h_n, a, s) + v(h_n, a, s)]$, and the *extremal sets* $B := (B_n)$ by $B_n(h_n) := \{a \in D_n(h_n)$: a maximizes $(L_n V_{n+1})(h_n, \cdot)\}$.

B is a *range* (since $|A| < \infty$). The following Theorem allows the description of the set of optimal plans by a range and will be basic for the minimization of variance in the set of expectation optimal plans. A similar result is [7, Corollary 1].

THEOREM 1.12. *Assume* $V > -\infty$.

(a) *If* $\pi \in \Delta$ *is optimal, we have* $\pi_n(h_n ; B_n(h_n)) = 1$ *for* $P_{n\pi}$*-almost all* $h_n \in H_n$, $n \in \mathbb{N}$.

(b) *If* C^+ *is fulfilled and* $\pi \in \Delta$ *has the property, that for all* $n \in \mathbb{N}$ *and* $P_{n\pi}$*—almost all* $h_n \in H_n$ *we have* $\pi_n(h_n ; B_n(h_n)) = 1$, *then* π *is optimal.*

Proof of (a). Let $\pi \in \Delta$ be optimal, and suppose there is a set C_n with $P_{n\pi}(C_n) > 0$, such that $\pi_n(h_n ; B_n(h_n)) < 1$ for all $h_n \in C_n$. Using the OE we obtain for all $h_n \in C_n$

$$\int_{B_n(h_n)} \pi_n(h_n; da)\,(L_n V_{n+1})\,(h_n, a) \leqslant \pi_n(h_n; B_n(h_n))\, V_n(h_n),$$

$$\int_{B_n(h_n)^c} \pi_n(h_n; da)\,(L_n V_{n+1})\,(h_n, a) < \pi_n(h_n; B_n(h_n)^c)\, V_n(h_n).$$

The sum yields a contradiction to the optimality of π, as from Lemma 1.10 (c) (for $n + 1$) we get

$$V_{n\pi}(h_n) = \int \pi_n(h_n; da)\,(L_n V_{n+1\pi})\,(h_n, a))$$

$$= \int \pi_n(h_n; da)\,(L_n V_{n+1})\,(h_n, a) < V_n(h_n) \qquad \text{for a.a. } h_n \in C_n,$$

and now look at Lemma 1.10 (c) (for n) again.

With use of C^+, we prove (b) in a constructive way: Let $\pi \in \Delta$ be as

described in part (b). For all $a \in B_n(h_n)$ we realize $V_n(h_n) = \sup_{b \in D_n(h_n)}(L_n V_{n+1})(h_n, b) = (L_n V_{n+1})(h_n, a)$, and thus

$$\begin{aligned} V_n(h_n) &= \pi_n(h_n; B_n(h_n))\,(L_n V_{n+1})\,(h_n, a) \\ &= \int_{B_n(h_n)} \pi_n(h_n; da)\,(L_n V_{n+1})\,(h_n, a) \\ &= \int \pi_n(h_n; da)\,(L_n V_{n+1})\,(h_n, a) \end{aligned}$$

for all $n \in \mathbb{N}$, $P_{n\pi}$-a.a. $h_n \in H_n$. This and $V_{n+1} \leqslant W_{n+1,+}$ imply $V_n(h_n) \leqslant (\Lambda_{n\pi}0)(h_n) + T_n W_{n+1,+}(h_n)$ for $P_{n\pi}$-a.a. $h_n \in H_n$, $n \in \mathbb{N}$, where we define for any integrable function v $(\Lambda_{n\pi}v)(h_n) := \int \pi_n(h_n; da) \int q_n(h_n, a; ds)\,[r_n(h_n, a, s) + v(h_n, a, s)]$. By downward induction we get

$$V_1(s) \leqslant (\Lambda_{1\pi}\Lambda_{2\pi} \cdots \Lambda_{n-1\pi}\Lambda_{n\pi}0)(s) + (T_{1n}W_{n+1,+})(s)$$

for q_0-a.a. $s \in S$, $n \in \mathbb{N}$, and as C^+ guarantees $\lim_k T_{1k}W_{1+k,+} = 0$, we have, by [1, Lemma 11.2] for q_0-a.a. $s \in S$ $V_1(s) \leqslant \lim_n (\Lambda_{1\pi}\Lambda_{2\pi} \cdots \Lambda_{n-1\pi}\Lambda_{n\pi}0)(s) = V_{1\pi}(s)$, and now Lemma 1.10 (b) completes the Proof.

DEFINITION 1.13. A DM is called *Markovian*, if D_n, q_n, r_n, $n \in \mathbb{N}$, depend on the history only by the present state, i.e., we have $D_n(s_n)$, $q_n(s_n, a_n; \cdot)$, $r_n(s_n, a_n, s_{n+1})$. Accordingly, we call a plan $\pi \in \Delta$ *Markovian*, if it is well defined by writing $\pi_n(s_n; \cdot)$, $h_n \in H_n$ with s_n as the last state of h_n, $n \in \mathbb{N}$. The set of Markovian plans out of Δ will be denoted as Δ_m.

LEMMA 1.14. ([1, Theorem 18.1, 18.4, 18.2]). *Let* DM *be Markovian.*

(a) *For a plan $\pi \in \Delta$ there is a $\sigma \in \Delta_m$ with $V_\sigma = V_\pi$.*

(b) $V_n = \sup_{\pi \in \Delta_m} V_{n\pi}$, $n \in \mathbb{N}$.

(c) *There is one and only one measurable extended real-valued map $V_n'\colon S \to \bar{\mathbb{R}}$ with $V_n(h_n) = V_n'(s_n)$ for $h_n \in H_n$, $n \in \mathbb{N}$, which fulfills the reduced OE, i.e., for all $s \in S$ we have*

$$V_n'(s) = \sup_{a \in D_n(s)} \int q_n(s, a; dt)\,[r_n(s, a, t) + V_{n+1}'(t)].$$

(d) *Let C^+ be fulfilled. A Markovian plan $\sigma \in \Delta_m$ is optimal, iff $\sigma_n(s_n; B_n'(s_n)) = 1$ for $P_{n\sigma}$-a.a. $h_n \in H_n$, s_n as the last state, with $B_n'(s) := \{a \in D_n(s)$: a maximizes $(L_n V_{n+1}')(s, \cdot)\}$, $n \in \mathbb{N}$.*

(e) *If C^+ is given, for any plan $\pi \in \Delta$ there is a deterministic Markovian plan $f \in \Delta_m$ with $V_f \geqslant V_\pi$.*

Part (d) and the finiteness of A assures the existence of an optimal plan, since for all $h_n \in H_n$ there is an $a_n \in B_n'(s_n)$, and the maps $f_n(h_n) := a_n$ define a deterministic plan out of Δ_m.

2. The Variance of the Random Reward

Till now we considered expectations of futural rewards. We are also interested in variances caused by expectation optimal plans. Let us construct a decision model DMV with identical state and action spaces and the same transition laws, but with the range of extremal sets and a reward function generating variances.

Here we use ideas comparable to Jaquette's paper [5] on stationary Markovian models. The Assumptions 1.7 were made to ensure the existence of the expectations $V_{n\pi}(h_n)$. Now we need assumptions for the second moments.

Assumption 2.1. *Presume C^+ and let Δ' be the set of optimal plans of* DM. *Assume,*

$$\sup_{\pi\in\Delta'} E_\pi[R_n^{\,2} \mid \chi_n = \cdot] < \infty, \qquad \lim_k T_{nk}(\sup_{\pi\in\Delta'} E_\pi[R_{n+k}^2 \mid \chi_{n+k} = \cdot]) = \ , \quad n \in \mathbb{N}.$$

These assumptions are fulfilled, if e.g., $\sum_{k=1}^{\infty} \| r_k \| < \infty$, where we understand $\| f \| := \sup_x |f(x)|$ for any function $f\colon X \to \mathbb{R}$. We define $Z_{n\pi}(h_n) := E_\pi[R_n^{\,2} \mid \chi_n = h_n]$, $h_n \in H_n$, $n \in \mathbb{N}$, $\pi \in \Delta'$. From $R_n^{\,2} = r_n^{\,2} + 2r_nR_{n+1} + R_{n+1}^2$ for $n \in \mathbb{N}$, we get for $h_n \in H_n$

$$Z_{n\pi}(h_n) = \int \pi_n(h_n; da) \int q_n(h_n, a; ds)\,[r_n^{\,2}(h_n, a, s) + 2r_n(h_n, a, s)\, V_{n+1\pi}(h_n, a, s) + Z_{n+1\pi}(h_n, a, s)].$$

For $\pi \in \Delta'$ we have $V_{n+1\pi} = V_{n+1}$ $P_{n+1\pi}$-a.s. Now we introduce the functions (b_n) by $b_n := r_n[r_n + 2V_{n+1}]$, $n \in \mathbb{N}$. Notice, that under C^+ r_n and V_{n+1} are bounded from above. For $n \in \mathbb{N}$, $\pi \in \Delta'$, $P_{n\pi}$-a.a. $h_n \in H_n$ we get

$$Z_{n\pi}(h_n) = E_\pi[b_n(\chi_{n+1}) + Z_{n+1\pi}(\chi_{n+1}) \mid \chi_n = h_n].$$

We continue this operation with downward induction:

$$Z_{n\pi}(h_n) = E_\pi\left[\sum_{m=n}^{n+k} b_m(\chi_{m+1}) + Z_{n+k+1\pi}(\chi_{n+k+1}) \mid \chi_n = h_n\right].$$

Using an induction and our assumption, we have

$$\lim_k E_\pi[Z_{n+k+1\pi} \mid \chi_n = h_n] \leqslant \lim_k T_{n,k+1} Z_{n+k+1\pi}$$

$$\leqslant \lim_k T_{nk} \sup_{\pi \in \Delta'} Z_{n+k\pi} = 0,$$

and so we get for $\pi \in \Delta'$, $n \in \mathbb{N}$, $P_{n\pi}$-a.a. $h_n \in H_n$:

$$0 \leqslant Z_{n\pi}(h_n) = \sum_{k=n}^{\infty} E_\pi[b_k(\chi_{k+1}) \mid \chi_n = h_n].$$

Thus the second moment is written as a sum of expectations.

Now we form the decision model DMV for the minimization of variance in Δ', and the symbols of DMV are marked by a "′".

DEFINITION 2.2. DMV is a DM with tuple

$$((S', \mathfrak{S}'), (A', \mathfrak{A}'), (D_n'), (q_n'), (r_n')) := ((S, \mathfrak{S}), (A, \mathfrak{A}), (B_n), (q_n), (-b_n)).$$

By Theorem 1.12, just the almost sure admissible plans of DMV are optimal in DM, and these plans form Δ' (cf. Lemma 1.4). In DMV we get $V' = \sup_{\pi \in \Delta'} V_\pi'$ by maximization of $V_\pi' = \int V_{1\pi}' \, dq_0$, and this is the minimization of $E_\pi[R^2] = \int Z_{1\pi} \, dq_0 = \sum_{k=1}^{\infty} E_\pi b_k(\chi_{k+1}) = -\int V_{1\pi}' \, dq_0 = -V_\pi'$ over Δ'. For all plans π of Δ' we have $\mathrm{Var}_\pi(R) = E_\pi(R^2) - (E_\pi R)^2$, and so we established dynamic programming of the random reward's variance.

If DM is Markovian, so is DMV (notice 1.14 (c), (d)). Thus in looking for minimum variance plans in the set of maximum expectation plans, we can restrict ourselves to the set of deterministic Markovian plans.

From $R_n^3 = r_n^3 + 3r_n R_{n+1}^2 + 3r_n^2 R_{n+1} + R_{n+1}^3$ we realize the straightforward continuation of this chapter: We have $V_{n\pi} = V_n$ and $Z_{n\pi} = Z_n$ for $P_{n\pi}$-a.a. arguments, $n \in \mathbb{N}$, $\pi \in \Delta''$ (the set of DMV-optimal plans π in Δ'), thus

$$E_\pi[R_n^3 \mid \chi_n = \cdot]$$

$$= E_\pi[r_n^3 + 3r_n Z_{n+1} + 3r_n^2 V_{n+1} + E_\pi[R_{n+1}^3 \mid \chi_{n+1} = \cdot] \mid \chi_n = \cdot].$$

With $r_n'' := (r_n^3 + 3r_n Z_{n+1} + 3r_n^2 V_{n+1})$ follows the formulation of the adequate model, which permits the maximization of the third moment in the set of maximum expectation, minimum variance plans, and so on.

3. The Semi-Markovian Decision Model SMDM

Let us introduce the well-known SMDM by the symbols $(J, (A_i), p_0, (p(i, a; j)), (F(i, a, j; \cdot)), (\hat{r}(i, a, j)), (\hat{s}(i, a, j)), \alpha)$. The finite set J consists of the states i, and for i is A_i the set of admissible actions. In state $i \in J$ under action $a \in A_i$, $p(i, a; j)$ is the probability for the next state being $j \in J$, and the probability that this transition will be finished during the course of s further units of time is $F(i, a, j; s)$.

When reaching j, the amount $\hat{r}(i, a, j)$ is paid. During the course of this transition the rate $\hat{s}(i, a, j)$ is gained per unit of duration. By continuous discounting, an amount B payable at time t has the present value (at time 0) $\exp(-\alpha t)B$, $\alpha > 0$ is the discount factor. The probability for the system starting in state i is given by $p_0(i)$. SMDM turns out to be a stationary Markovian DM, if we use the following technical definition:

Definition 3.1. *of* SMDM

(a) $(S, \mathfrak{S}) := (\mathbb{R} \times J, \mathfrak{B} \otimes \mathfrak{P}(J))$, where $\mathfrak{B}$ is the usual Borel field. $s_n = (t, i) \in S$ symbolizes, that at time t from the beginning the state $i \in J$ is present.

(b) $A := \bigcup_{i \in J} A_i$, $|A_i| < \infty$, $i \in J$.

(c) $D_n(h_n) := A_i$ for $h_n \in H_n$ with $s_n = (t, i)$.

(d) $q_0(0, i) := p_0(i)$, $i \in J$.

$q_n(h_n, a; \cdot) := q(t_n, i_n, a; \cdot, \cdot)$ for $h_n \in H_n$ with $s_n = (t_n, i_n)$, $a \in A_{i_n}$, with $q(t, i, a; B, j) := p(i, a; j) \int_{(B-t)} F(i, a, j; dx)$, $i \in J$, $B \in \mathfrak{B}$, where we use $(B - t) := \{x \in \mathbb{R}: x + t \in B\}$. We can write t_n in terms of the inter state transition times x_k, and get $t_n = \sum_{k=1}^{n} x_k$.

(e) $r_n(h_n, a, s_{n+1}) := \exp(-\alpha t)\, s(i, a, x, j)$, with $s(i, a, x, j) := \hat{r}(i, a, j) \times \exp(-\alpha x) + \hat{s}(i, a, j)\, 1/\alpha(1 - \exp(-\alpha x))$, $n \in \mathbb{N}$, $h_n \in H_n$ with $s_n = (t, i)$ and $s_{n+1} = (t + x, j)$. For convenience, we introduce $g(i, a) := \sum_J p(i, a; j) \times \int F(i, a, j; dx)\, s(i, a, x, j)$. To get the (natural) condition C^+ fulfilled, we presume:

Assumption 3.2.

(a) $\|\hat{r}\| \leqslant c_1 < \infty$, $\|\hat{s}\| \leqslant c_2 < \infty$.

(b) $0 < \int x F(i, a, j; dx) < \infty$, $i, j \in J$, $a \in A_i$.

As an immediate consequence follows for the distributions $\hat{F}(\cdot) := \text{minimum}_{i, a \in A_i, j} F(i, a, j; \cdot)$ and $F(\cdot) := \text{maximum}_{i, a \in A_i, j} F(i, a, j; \cdot)$ and for the expected discountation for a transition from i to j under a, $\alpha(i, a, j) := \int \exp(-\alpha x) F(i, a, j; dx)$:

LEMMA 3.3. (a) $0 < \int_0^\infty xF(dx) \leqslant \int_0^\infty x\hat{F}(dx) < \infty$.

(b) $|g(i, a)| \leqslant c_1 + c_2 \int_0^\infty x\hat{F}(dx) < \infty$, $i \in J$, $a \in A_i$.

(c) $1 > \beta := \int \exp(-\alpha x) F(dx) \geqslant \alpha(i, a, j) \geqslant \int \exp(-\alpha x)\hat{F}(dx) > 0$.

PROPOSITION 3.4.

(a) *$E_\pi R$ exists for all $\pi \in \Delta$, and*

$$K := \left(c_1 + c_2 \int x\hat{F}(dx)\right) \sum_{k=0}^{\infty} \beta^k \geqslant |E_\pi R|.$$

(b) $\sup_{\pi \in \Delta} E_\pi |R| \leqslant K$ *and* $V_n = \sup_{\pi \in \Delta} E_\pi[R_n \mid \chi_n = \cdot] \leqslant K$, $n \in \mathbb{N}$.

(c) SMDM *fulfills condition* C^+.

Using the Theorem of Ionescu–Tulcea and Lemma 3.3, the proof of (a) is a matter of straightforward calculations, (b), (c) are immediate consequences of (a) and

$$\lim_k \| T_{nk} W_{n+k,+} \| \leqslant \lim_k \| T_{nk} K \| \leqslant \lim_k \left(\int \exp(-\alpha x) F(dx)\right)^k K = 0 .$$

Now we deduce the following results from Lemma 1.14.

THEOREM 3.5.

(a) *For a plan $\pi \in \Delta$, there is a deterministic Markovian plan $f \in \Delta_m$ with $V_f \geqslant V_\pi$.*

(b) *For $n \in \mathbb{N}$, V_n depends only on the s_n-part of h_n, and is written $V_n(t_n, j_n)$ for $s_n = (t_n, j_n)$.*

(c) *For $n \in \mathbb{N}$, $t \in \mathbb{R}^+$, $i \in J$ we have $V_n(t, i) = \exp(-\alpha t)\, V_1(0, i)$.*

(d) *With $G(i) := V_1(0, i)$, $i \in J$, the OE is reduced to the form $G(i) = \max_{a \in A_i} \{g(i, a) + \sum_J p(i, a; j)\, \alpha(i, a, j)\, G(j)\}$, $i \in J$.*

(e) *For $n \in \mathbb{N}$, $h_n \in H_n$ with $s_n = (t_n, i_n)$, the extremal sets are $B_n(h_n) = B(i_n)$, with $B(i) := \{a \in A_i : g(i, a) + \sum_J p(i, a; j)\, \alpha(i, a, j)\, G(j)$ is maximal in $A_i\}$. There exists an optimal deterministic stationary Markovian plan $f = (f_n) \in \Delta$, which only depends on the last state out of J. This optimal plan has the form $f_n(h_n) = f(j_n)$, $n \in \mathbb{N}$, $h_n \in H_n$ with $s_n = (t_n, j_n)$. The set of plans of this kind is finite.*

Proof. For parts (a) and (b) see 1.14 parts (e) and (c). To show part (c) we admit Markovian plans only, in accordance with 1.14 (a). For $\pi \in \Delta_m$, $n \in \mathbb{N}$, $h_n \in H_n$ with $s_n = (t, j)$, we have

$$V_{\pi n}(h_n) =$$

$$\sum_{k=n}^{\infty} \exp(-\alpha t) \sum \pi_n(t, j; a_n) \sum p(j, a_n; j_{n+1}) \int F(j, a_n, j_{n+1}; dx_{n+1})$$

$$\exp(-\alpha x_{n+1}) \cdots \sum p(j_{k-1}, a_{k-1}; j_k) \int F(j_{k-1}, a_{k-1}, j_k; dx_k)$$

$$\exp(-\alpha x_k) \sum \pi_k \left(t + \sum_{m=n+1}^{k} x_m, j_k; a_k \right) g(j_k, a_k).$$

So we can write $V_{n\pi}(h_n)$ as a sum of terms, which depend on h_n only by $j_n = j$ and $t_n = t$, and get by taking the supremum on both sides $V_n(h_n) = \exp(-\alpha t_n)\, V_1(0, j_n)$. Part (d): We have 1.14 (c) in the form

$$V_n(t, i) = \max_{a \in A_i} \left\{ \exp(-\alpha t)\, g(i, a) + \sum p(i, a; j) \int F(i, a, j; ds)\, V_{n+1}(t + s, j) \right\},$$

and by part (c) above, the desired equation is obvious. Part (e): From the form in the proof of (d), we at once see the structure of the extremal sets. C^+ is fulfilled. Due to 1.14 (d) a Markovian plan σ is optimal, iff

$$\sigma_n(t_n, i_n; B(i_n)) = 1 \qquad \text{for} \quad P_{n\sigma}\text{-a.a. } h_n \in H_n .$$

As $|A_i| \neq 0$, $i \in J$, we have $B(i) \neq \varnothing$, $i \in J$, So let us define $f : J \to A$, $i \to f(i) := a_i \in B(i)$, $i \in J$. By the theorem mentioned above, f is optimal.

Notice, that f is independent from p_0 ; $f := (f, f, \ldots)$ is a *stationary*, so called *strongly optimal* plan.

Our task is to find an optimal plan for SMDM. We are justified to restrict ourselves to the finite set F of maps $f : J \to A$, $f(i) \in A_i$, $i \in J$. The Proof of part (e) was nonconstructive, as in general we do not know the $B(i)$, which are determined by $G(j) = \sup_{\pi \in \Delta} V_{1\pi}(0, j)$. Osaki and Mine [8] suggested the construction of stationary Markovian optimal plans via linear programming. Their method even work in our concept, which is more general regarding the structure of the SMDM. We sketch the method without Proofs, which in detail can be seen in [9]. Using the Kronecker symbol δ, the linear program LP is:

$$\sum_{j \in J} \sum_{a \in A_j} g(j, a)\, x(j, a) = \max_x, \qquad x(j, a) \geqslant 0, \quad j \in J, \quad a \in A_j$$

$$\sum_{j \in J} \sum_{a \in A_j} [\delta_{jk} - \alpha(j, a, k)\, p(j, a; k)]\, x(j, a) = p_0(k), \qquad k \in J.$$

It's usability is a consequence of

THEOREM 3.6.

(a) *There exist extreme solutions* x *of* LP, *that for* $j \in J$ *either* $x(j, a) = 0$ *for all* $a \in A_j$, *or there is exact one* $a_j \in A_j$ *with* $x(j, a_j) > 0$ *and* $x(j, a) = 0$ *for all* $a \neq a_j$. *Let* $I(x) := \{j \in J :$ *there is an* $a_j \in A_j$ *with* $x(j, a_j) > 0\}$.

(b) *To any extreme solution* x *of the form given in* (a), *we define* $f \in F$ *by* $f(j) := a_j$, $j \in I$, *and* $f(j) \in A_j$ *arbitrarily for* $j \in J \setminus I$. *This* f *is optimal for* SMDM.

Thus linear programming establishes a solution of SMDM. Another well known method is the *Howard iteration*, also denoted as *plan iteration*: An arbitrary $f_0 \in F$ is improved to f_1, where we define f_1 by the maximum points a_i, $i \in J$, of the map $a \to g(i, a) + \sum p(i, a; j)\, \alpha(i, a, j)\, G_{f_0}(j)$, which is defined on A_i, $i \in J$, with $G_f(i) := V_{1f}(0, i)$. The Howard iteration may be accelerated by a preceding *value iteration*: Starting from $G_0(i) := 0$, $i \in J$, we compute $G_n(i) := \max_{a \in A_i} \{g(i, a) + \sum p(i, a; j)\, \alpha(i, a, j)\, G_{n-1}(j)\}$, $i \in J$. If G_n is "almost equal" G_{n-1}, we take the function of these maximum points as f_0. Till now we saved solving the solution of linear equation systems in G_f and can hope to have a good initial f_0. The convergence of G_n to G can be verified (cf. Schäl [11] for Theorems about "Wertiteration" and "Planiteration").

4. COMPUTING THE VARIANCE OF SMDM

According to Definitions 2.2 and 3.1, we introduce SMDMV.

DEFINITION 4.1. SMDMV is a DM with tupel

$$((S', \mathfrak{S}'), (A', \mathfrak{A}'), (D_n'), (q_n'), (r_n')) :=$$

$$\left((\mathbb{R} \times J, \mathfrak{B} \otimes \mathfrak{P}(J)), \left(\bigcup_{i \in J} A_i, \mathfrak{P}(A')\right), (B_n), (q_n), (-b_n)\right),$$

with $b_n := r_n(r_n + 2V_{n+1})$, where the A_i, B_n, q_n, r_n and V_n are as in Section 3, especially $B_n(h_n) = B(j_n)$. For $s_n = (t, i)$ and $a_n = a$, $s_{n+1} = (t + x, j)$ we have from 3.5 (c), (d):

$$\begin{aligned} b_n(h_n, a_n, s_{n+1}) &= -\exp(-2\alpha t)\, s'(i, a, x, j) \\ &= \exp(-2\alpha t)\, s(i, a, x, j)\, [s(i, a, x, j) + 2 \exp(-\alpha x)\, G(j)]. \end{aligned}$$

Thus we realize SMDMV to be a stationary Markovian decision model, of the same type as SMDM with $A_i' = B(i)$ and $\alpha' = 2\alpha$. For convenience again, we introduce $b(i, a) := -\sum_J p(i, a; j) \int F(i, a, j; dx)\, s'(i, a, x, j)$, and

define $\beta(i, a, j) := \int \exp(-2\alpha x) F(i, a, j; dx)$. By separate integration of the parts of $b(i, a)$, we realize $b(\cdot, \cdot)$ to be bounded. To continue as in Chapter 2, we have to make sure Assumption 2.1.

THEOREM 4.2. $\sup_{\pi \in \Delta'} E_\pi[R^2] < \infty$, *Assumption* 2.1 *is fulfilled.*

Proof. From Assumption 3.2 we get

$$E_\pi R^2 = \int \left(\sum_{k=1}^{\infty} \exp(-\alpha t_k)\, s(j_k, a_k, x_{k+1}, j_{k+1}) \right)^2 dQ_\pi$$

$$\leqslant (c_1 + c_2/\alpha)^2 \int \left(\sum_{k=1}^{\infty} \exp(-\alpha t_k) \right)^2 dQ_\pi, \qquad \pi \in \Delta.$$

With $t_k = \sum_{u=1}^{k} x_u$ we try to give bounds for

$$E_\pi \left(\sum_{k=1}^{\infty} \exp\left(-\alpha \sum_{u=1}^{k} x_u \right) \right)^2$$

$$= E_\pi \left(\sum_{k=1}^{\infty} \left[\left(\prod_{u=1}^{k} \exp(-\alpha x_u) \right)^2 \left(1 + 2 \sum_{m=k+1}^{\infty} \prod_{v=k+1}^{m} \exp(-\alpha x_v) \right) \right] \right).$$

The Theorem of Ionescu–Tulcea shows the independence of the x_k for a given history of states and actions. So the expectation above is equal to

$$E_\pi \left(\sum_{k=1}^{\infty} \left[\left(\prod_{u=1}^{k} \exp(-2\alpha x_u) \right) \left(1 + 2E_\pi \left(\sum_{m=k+1}^{\infty} \left[\prod_{v=k+1}^{m} \exp(-\alpha x_v) \right] \Big| \chi_k \right) \right) \right] \right).$$

The expectation on the right can be regarded as the profit in a SMDM with $\hat{r} \equiv 1$, $\hat{s} \equiv 0$ and discountation α, so by 3.4, it is bounded by a constant. The total expectation can be estimated by this constant and the expected profit of a SMDM with $\hat{r} \equiv 1$, $\hat{s} \equiv 0$ and discountation 2α, which is bounded by another constant. The result is $E_\pi R^2 < \infty$, $\pi \in \Delta$. In the same way we get $E_\pi[R^2 \mid \chi_1 = i] < \infty$, $i \in J$. As in Proposition 3.4 (c) we realize, that in SMDMV condition C^+ is fulfilled (remember the similarity of structure). Now from the estimation above, we have

$$\sup_{\pi \in \Delta'} E_\pi Z_{n\pi} = \sup_{\pi \in \Delta'} E_\pi(E_\pi[R_n^2 \mid \chi_n = \cdot]) \leqslant \sup_{\pi \in \Delta'} E_\pi(\sup_{\pi \in \Delta'} E_\pi[R_1^2 \mid s_1 = \cdot]) < \infty,$$

and realize $\lim_k T_{nk} \sup_{\pi \in \Delta'} E_\pi[(R_{n+k})^2 \mid \chi_{n+k} = \cdot] = 0$, too.

COROLLARY 4.3.

(a) *In* SMDMV *we have* $V_n'(h_n) = \exp(-2\alpha t_n)\ G'(j_n)$ *for* $n \in \mathbb{N}$;

$h_n \in H_n$ with $s_n = (t_n, j_n)$, where $G'(i) := V_1'(0, i)$, $i \in J$, and the reduced optimality equation is

$$G'(i) = \max_{a \in B(i)} \left\{ -b(i, a) + \sum_J p(i, a; j)\, \beta(i, a, j)\, G'(j) \right\}.$$

(b) *In* SMDM *exists a deterministic stationary Markovian plan, denoted by $f \in F$, which is expectation optimal with minimum variance.*

(c) *$G_f'(i) := V_{1f}'(0, i) = \exp(2\alpha t)\, V_{nf}'(h_n)$, for $n \in \mathbb{N}$, $f \in F$, $h_n \in H_n$ with $s_n = (t, i)$, fulfills the system*

$$G_f'(i) = -b(i, f(i)) + \sum_J p(i, f(i); j)\, \beta(i, f(i), j)\, G_f'(j) \qquad i \in J.$$

This system of linear equations has an unique solution, and the variance to f is $\mathrm{Var}_f(R) = -\sum_J p_0(i)\, G_f'(i) - (E_f R)^2$.

(d) *The decision function f in* (b) *can be obtained by solving two linear programs, or by processing two Howard (value) iterations.*

As Assumption 2.1 is fulfilled, Corollary 4.3 (a), (b) is nothing but Theorem 3.5 applied to SMDMV. We just have to watch the notational changes and the relationships between the models SMDM and SMDMV, such as for $n \in \mathbb{N}$, $\pi \in \Delta'$ $-V_{n\pi}' = E_\pi [\sum_{k=n}^{\infty} \exp(-2\alpha t_k)\ b(j_k, a_k) | \chi_n = \cdot] = Z_{n\pi}$ $P_{n\pi}$-a.s., and $\mathrm{Var}_\pi(R) = \int Z_{1\pi}\, dq_0 - V_\pi{}^2$. Parts (c) and (d) are analogous to Theorem 3.6, and the linear program for this problem is LP′:

$$\sum_{j \in J} \sum_{a \in B(j)} -b(j, a)\, x(j, a) = \max_x, \qquad x(j, a) \geqslant 0, \quad j \in J, \quad a \in B(j),$$

$$\sum_{j \in J} \sum_{a \in B(j)} [\delta_{jk} - \beta(j, a, k)\, p(j, a; k)]\, x(j, a) = p_0(k), \qquad k \in J.$$

Acknowledgments

We express our acknowledgment to the referees, whose detailed comments and suggestions improved the comprehensibility of the paper to a great deal.

References

1. K. Hinderer, Foundations of non-stationary dynamic programming with discrete time parameter, *in* "Lecture Notes in Operations Research and Mathematical Systems," Springer–Verlag, Berlin, 1970.
2. K. Hinderer, Instationäre dynamische Optimierung bei schwachen Voraussetzungen über die Gewinnfunktionen, *Abh. Math. Sem. Univ. Hamburg* **36** (1971), 208–223.

3. R. A. HOWARD, Research in semi-Markovian decision structures, *J. Operations Res. Soc. Japan* **6** (1964), 163–199.
4. S. C. JAQUETTE, Markov decision processes with a new optimality criterion: small interest rates, *Ann. Math. Statist.* **43** (1972), 1894–1901.
5. S. C. JAQUETTE, Markov decision processes with a new optimality criterion: discrete time, *Ann. Statist.* **1** (1973), 496–505.
6. W. S. JEWELL, Markov renewal programming I, II, *Operations Res.* **11** (1963), 938–948, 949–971.
7. P. MANDL, On the Variance in controlled Markov chains, *Kybernetika* **7** (1971), 1–12.
8. S. OSAKI AND H. MINE, Linear programming algorithms for semi-Markovian decision processes, *J. Math. Anal. Appl.* **22** (1968), 356–381.
9. G. QUELLE, Dynamische Optimierung Semi-Markoffscher Entscheidungsmodelle: Stationaritätseigenschaften, Varianzen, präventive Aktionen, Doctoral Dissertation, Universität Hamburg, 1973.
10. U. RIEDER, Bayessche dynamische Entscheidungs- und Stoppmodelle, Doctoral Dissertation, Universität Hamburg, 1972.
11. M. SCHÄL, Dynamische Optimierung unter Stetigkeits- und Kompaktheitsbedingungen, Habilitationsschrift, Universität Hamburg, 1972.

Printed in Belgium